特约专稿
DOI：10.13998/j.cnki.issn1002-1248.20-0969

情报认知模型库构建研究

刘细文[1,2]，郭世杰[1,2]
（1. 中国科学院大学 经济与管理学院图书情报与档案管理系，北京 100049；2. 中国科学院文献情报中心，北京 100190）

摘　要：［目的／意义］研究情报认知模型库的组织、建设方法，帮助科研人员和情报分析人员快速形成对学科领域的准确认知，协助识别科技机遇、甄别技术威胁。［方法／过程］情报认知模型库包含跨领域的各种技术要素和文献信息，因此在结构上包括文献库、算法库、学科领域知识库、研究应用案例库等，支持研究热点识别、技术性能对比、智能分析方法推荐、算法辅助设计等功能。模型库的建设流程包括对情报认知模型的收集、验证、存储、组织、利用等，其中对模型的验证是至关重要的一步。［结果／结论］构建情报认知模型库对开展科学技术领域情报工作具有重要意义，能够发挥科技数据基础设施、科技情报分析工具箱的效果。情报认知模型库的构建需要情报人员、科学技术领域专家、信息技术人员的通力合作，未来需要进一步考虑模型库的维护升级、推广应用等问题。

关键词：情报认知模型；模型库；情报研究基础设施；学术情报研究
中图分类号：G350；G255.51　　**文献标识码：**A　　**文章编号：**1002-1248（2021）01-0032-09

引用本文：刘细文, 郭世杰. 情报认知模型库构建研究[J]. 农业图书情报学报, 2021, 33(1)：32-40.

A Database Construction of S&T Intelligence Cognition Models

LIU Xiwen[1,2], GUO Shijie[1,2]
(1. Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049; 2. National Science Library, Chinese Academy of Sciences, Beijing 100190)

Abstract: [Purpose/Significance] This paper aims to study the organization and construction methods of the intelligence cognition models database to help scientists and information analysts to have an accurate understanding of the research area within a short period of time, and assist them in identifying technological opportunities and threats. [Method/Process] The intelligence cognition models database contains various technical elements and literature

收稿日期：2020-10-28
基金项目：中国科学院文献情报能力建设专项课题“科技知识服务大数据基础设施”的子课题“技术领域情报挖掘模型”（2020WQZX0001）；中国科学院文献情报中心青年人才创新团队项目“面向科研设施的信息抽取”（G180141）
作者简介：刘细文（ORCID：0000-0003-0820-3622），男，博士，二级研究员，博士生导师，研究方向为情报学。郭世杰（ORCID：0000-0003-2782-7553），男，博士在读，助理研究员，研究方向为情报学理论与方法、空间光电科技情报

information across different subjects, so it consists of literature library, algorithm library, scientific/technical knowledge library, application cases library, etc. It has functions including research hotspots identification, technical performance comparison, information analysis methods recommendation, algorithm-aided design and so on. The construction process of the database includes the collection, verification, storage, organization, and utilization of the "intelligence cognitive models", among which the verification of the models is a crucial step. [Results/Conclusions] The intelligence cognition models database is of great significance to the scientific and technological information study and it can play the role of data infrastructures and information analysis toolboxes. The construction of the database requires the cooperation of the information analysts, scientists, and information technology specialists. In the future, the maintenance, application and upgrading of the models library need to be further considered.

Keywords: intelligence cognition model; models database; information research infrastructure; academic information research

1 引　言

现代科学研究在微观、宏观、复杂性等方面不断深入，“数据密集型科学发现”正日益发挥不可替代的作用，多学科交叉前沿和一系列颠覆性技术正在不断塑造新的科学研究竞争格局；面对不断涌现的新技术、新知识、新概念，如何与前沿科学家和工程师同步认知最新科技发展态势，高效和准确地还原不同研究主题和领域知识本体全貌，进而从纷繁复杂的技术领域动态中敏捷地监测、抓取、挖掘出有效信息，为决策者和科研人员搭建好“从数据到信息，从知识到智慧”的桥梁，是情报工作者面临的重要问题。

科研数据的爆炸式增长对大规模知识管理和基于机器理解的知识挖掘带来了新的挑战，同时也提供了对跨学科知识进行集成和解析挖掘的可行条件。2012年，TONY 等在《第四范式：数据密集型科学发现》[1]中指出，未来的计算机系统应当能够自动发现、获取、组织、分析、关联、解释、推断信息，在全球范围内管理和处理知识的基础架构可作为下一代知识驱动型服务和应用程序的基础，研究人员可以利用这样的基础设施，提出与他们专业领域相关的问题，并在这样的“知识海洋”中找到答案。2020 年，张霖[2]提出，通过利用物理模型、传感器更新、运行历史数据等资源，可以集成多学科、多物理量、多尺度、多概率的仿真，可以在虚拟空间中完成对现实物理对象的映射，从而进行分析、预测、诊断、训练等，开展学科领域的知识发现研究。

在人工智能技术快速发展的背景下，构建不同科技领域的“认知模型”，利用算法和计算框架对物理世界的各种研究对象进行抽象和描述，是进一步利用各种计算工具、服务和应用程序，实现机器辅助知识推理、演绎、跨领域相关分析的基础。近年来，国内外已有大量机器学习算法将科学研究的问题抽象为包含该领域核心知识的若干关键特征，这些特征包括技术性能参数、结构成分、材料、制备方法、生产工艺、应用方向等，为快速构建对该学科领域的情报认知、帮助一线科研人员积累科学研究方法和思路、指导情报研究人员理解特定学科领域的知识背景等提供了便利条件。

2 学科领域的知识挖掘探索与实践

随着科研数据的不断丰富和信息技术的高速发展，国内外许多研究人员构建了专业学科领域的分析模型，在此基础上利用公开发表的文献资料，对该领域的关键知识与信息（如新材料、新工艺、新研究方向等）进行挖掘和分析。2019 年 7 月，TSHITOYAN 等在 *Nature* 上报道了利用无监督词嵌入模型从材料科学文献中发现潜在新材料组份的方法[3]，利用近 330 万份论

DOI：10.13998/j.cnki.issn1002-1248.20-0969

文摘要数据训练词嵌入模型，对文献中材料的“结构–属性”关系进行了挖掘，预测了可能具有较高热电品质因数的新材料，并基于历史数据成功开展回溯测试验证，表明词向量空间模型的位置编码可能包含材料科学知识。2015 年，ROSS 利用“机器人科学家”开展化学和生物研究的方法，构建了包含 1 万多个知识点的 11 层深的嵌套树状结构模型，将实验假设、测量结果、目标等知识以具有相关概率的逻辑进行表示并通过语义网进行发布，发现了对抗热带病的新铅化合物[4]。2019 年 7 月，FATHALLA 等报道了“科学事件数据模型”（OR-SEO）构建和应用方法，对人员、组织、位置、时间等科学事件“要素”及它们之间的关系进行建模，并通过组合规则发现要素间新的关系、推断知识图谱中未明确的新知识[5]。与此类似的还包括许多用于分析和发现新基因、新药物的生物信息学（Bio-Informatics）、医学信息学（Medical Informatics）挖掘模型等。

在更广泛的科学技术领域，应用各种机器学习方法直接从各种实验、观测、分析、测量数据中发掘新知识的研究大放异彩，取得了令人瞩目的丰富成果，近年来在人工智能的热潮中已经广为人知。例如，在生物学中，KOOHI-MOGHADAM 等利用多通道卷积神经网络（MCCNN）模型，对医学数据库中 11 万余条致病蛋白质突变数据、16 万余条金属结合位点的蛋白质三维结构数据进行了分析，揭示了十几种疾病和不同金属、不同蛋白质突变之间的相关关系[6]；在天文学中，DATTILO 等利用多个卷积神经网络对“开普勒”（Kepler）空间望远镜的观测数据进行分析，在距地球 1 200 光年的水瓶座星座中发现了 2 颗系外行星[7]；在太阳物理学中，WANG 等利用核主成分分析（KPCA）模型对太阳耀斑先兆因子（磁通量、磁螺度平均值等）数据进行了分析，增强了对强太阳耀斑的预报能力[8]；在地质学中，PHAM 等利用决策树（DT）分类器、基于旋转森林的决策树（RFDT）、基于多重提升的决策树（MDT）等模型，对印度某地区的 10 种地质数据（海拔、降水量、坡度、河流密度、岩性、地形湿度指数等）进行了分析，完成了对 34 口地下水井水位的预测[9]。

上述研究均构建了用于分析某一学科领域数据的学科认知模型，这些模型包含对相关科学/技术工程的关键问题、核心技术、性能指标、相关关系的描述和抽象，提供了研究这些科学/技术工程问题的方法、思路、计算框架、核心算法等，代表着科研人员对特定学科领域的认知。若能将这些模型进行有效解析、存储，实现可查询、检索、重复使用，将为不同领域的科研人员和情报分析人员提供快速切入最新研究领域、敏捷获取领域知识本体的抓手。

3 建设情报认知模型库的必要性与意义

3.1 什么是情报认知模型？

情报是知识的流动，并可以带来原有知识结构的变化。科学进步依赖于对现有知识的有效吸收，以选择最有前途的演进方向发展，并最大程度地减少重复劳动。潜在知识一方面蕴藏在科学研究实验/测量/观测/分析数据中，另一方面也很大程度上“沉淀”在已有的学术文献中。如果通过从大量学术文献中提取知识和关系，能够揭示“沉淀”知识，带来全新的开发和设计成果，使原有知识结构发生改变、形成新的知识结构。正如布鲁克斯知识方程[10]描述的那样：

$$K[S]+\Delta I = K[S+\Delta S] \quad (1)$$

方程（1）中，$K[S]$代表原有知识结构，ΔI为情报增量，$K[S+\Delta S]$为新的知识结构。这里的ΔI既可以来自各种自然科学实验、观测、测量、计算活动，也可以来自对文献资源和各种音视频媒体信息的综合、提炼、归纳、总结、对比、分析，而后者正是学科情报工作的基本内涵之一。

基于已有学术文献，开展深度情报研究、快速应对技术威胁、准确把握学科研究进展等，都需要高效地对学科领域与技术优势形成快速准确科学认知。然而，这种科学认知除了需要借助信息化、智能化手段综合分析科学大数据、学术文献大数据等外，还需要借助各个学科领域的认知模型，通过信息计算、数据

计算、情报计算方式形成对学科和技术态势的情报认知。如果广泛收集学科知识认知模型建立“情报认知模型库”，则可以作为新的情报研究数据基础设施，帮助不同领域的科研工作者、情报人员快速识别新科技理念、发现科技机遇、甄别技术威胁，进行知识挖掘、组织、集成、关联、重组。

基于以上认识和发展需要，可以将“情报认知模型”定义为：科研人员借助于文献信息资源、实验数据进行分析和挖掘的模拟、仿真计算模型，以及相关的知识挖掘计算方法等。

3.2 “情报认知模型”的类型与作用

在不同学科领域，已有许多研究开发和构建了各种分析模型，如前文提到的材料性能挖掘模型、蛋白质结构模型、系外行星识别模型、太阳耀斑活动分析模型、地下水水位分析模型等，可以将它们看作相关学科领域的“情报认知模型”。但是，情报研究人员还很少从特色数据资源和数据基础设施的角度去看待这些“情报认知模型”，也很少从工程化实施的角度，去建设一个解析、存储、检索和调用这些“情报认知模型”的信息库。

以石墨烯材料领域为例，相关研究可涉及至少 5 类信息对象：第 1 类是开展实验分析或测量所获得的科学实验数据，如进行石墨烯导电性实验时记录的时间信息和各种实验仪器读数等；第 2 类是从这些实验数据中获得的知识，例如可以是石墨烯的电学性能、热学性能、光学性能、力学性能、制备方法、功能化应用领域等；第 3 类是对科学技术领域的知识进行挖掘、组织、归纳、分析之后所获得的情报。例如可以是石墨烯的各种性能参数和应用领域、制备方法之间的因果关系、上下位关系、包含关系、“材料—成品”关系、“实体—值”关系等。第 4 类是用于指导对科学技术领域知识进行分析和挖掘的认知模型，例如可以是抽取石墨烯的热学、力学、电学性能特征、采用聚类算法分析石墨烯研究主题的无监督学习模型，也可以是抽取石墨烯的制备方法和催化剂种类特征的机器学习模型等。第 5 类则是对各种情报认知模型进行分类、解析、组织和结构化存储的模型库，例如不同石墨烯情报认知模型的训练数据源、特征抽取规则、验证方法、情报挖掘和预测效果、源代码等。上述 5 类信息对象都对科学技术领域的科学研究和科技情报工作具有价值，都可以进行积累和收集，对科研人员和情报人员提供服务，发挥“数据基础设施”和“情报分析/科学研究工具箱”的功效。

3.3 国内外已建成的公开数据基础设施尚未覆盖“情报认知模型库”功能

当前，许多研究领域呈现出对长期连续观测获取数据、有效存储和传输数据、多源数据综合分析等能力的强烈需求。在这样的趋势下，国内外已经建设了许多“数据基础设施”，这其中包括美国能源部于 2019 年部署的“环境科学虚拟生态系统数据基础设施”（ESS-DVIE）[11]、法国于 2018 年规划升级的“法国国家核物理和粒子物理计算中心”（CC-IN2P3）[12]、日本国立遗传学研究所（NIG）建设的“日本 DNA 数据银行”（DDBJ）[13]、欧洲将于 2021 年建成“多尺度植物表型组学和模拟欧洲设施”（EMPHASIS）和已建成并运行的“欧洲生物信息分布式网络”(ELIXIR)[14]等。在中国，科技部和财政部于 2019 年 6 月对国家科技资源共享服务平台进行了整合，形成了“国家高能物理科学数据中心”等 20 个国家科学数据中心、“国家重要野生植物种质资源库”等 30 个国家生物种质与实验材料资源库[15]。此外，中国科学院计算机网络信息中心建设了“中国科学院数据云”[16]，中国科学院文献情报中心建设了“科技文献大数据知识资源体系”[17]等。

与上述已建成的“数据基础设施”相比，“情报认知模型库”最大的不同在于其囊括了诸多独特的应用场景、科学技术领域知识，以及对各种科研问题的抽象方法和仿真框架。不仅如此，这些场景、知识、方法、框架是相互关联的，能够被“模型库”的用户统一检索、查询、调用。而作为对比，前文提到的大部分现有“数据基础设施”只存储了科学实验研究的底层实验数据，或存储了从底层实验数据中提炼、总

DOI：10.13998/j.cnki.issn1002-1248.20-0969

结发现的学科知识；尽管少数“数据基础设施”对一些通用的数据挖掘算法进行了存储，但是这些算法是孤立的，并不包含细分科学技术领域的知识框架、特征抽取标准，因此难以与这些领域的具体应用场景进行关联。因此，尽管“情报认知模型库”在数据来源和存储内容上与现有的数据基础设施具有一定联系和相似性，但是它们之间依然存在显而易见的差异。

此外，通过对模型库中各种跨科学技术领域的、从数据到知识的分析模型进行对比、归纳、总结，“模型库”构建人员未来将可能就“人工智能技术赋能的数据挖掘和知识发现”的共性方法进行更深入的分析和研究，从而在数据密集型知识发现的研究范式、理论、方法上做出更多贡献。

3.4 “情报认知模型库”可以为甄别科技机遇和威胁提供快速支撑

科学技术的最新进展往往带来全新的科技认知，可能是新关注焦点、新应用场景、新发展方向，也可能是性能的突破、方法工艺的革新、结构成分的突变等。这些新的科技认知不断冲击旧的知识体系，在原有知识结构中催生了新的知识节点、形成了新的关联关系，亦或者突破了原有认知模型中所存储的参数阈值。为了对这些最新科技进展进行有效甄别，必须保证作为本底信息的原有知识储备的全面性、准确性、专业性。

以包含“情报认知模型”的高水平学术论文为收集对象，通过制定对各种科学技术领域情报认知模型的标引、融合、验证、更新规则，构建跨领域情报认知体系的规范和框架，有利于积累相关数据和模型方法，发展面向关键核心技术性能评价的指标，乃至通过对各种认知模型的集成、关联、重组、梳理，逐渐形成跨领域、大规模、结构化的科学技术领域知识库。届时，根据技术性能参数比对、成分结构查询、技术工艺对比等方法，就能够快速甄别新科技动态中所蕴含的机遇和威胁，或利用综合性指标体系判断新科技成果的突破和创新颠覆程度。

4 “情报认知模型库”的建设方法设计

正是由于“情报认知模型库”的重要意义和价值，十分有必要厘清模型库的结构和功能，设计一套行之有效的模型库构建方法。

4.1 “科学技术领域情报认知模型”的要素和关系

一般而言，学科领域知识挖掘模型是由各领域科研人员开发，并以学术论文的形式进行报道和公开的。在许多情况下，这些模型会将某一类研究对象或科学问题抽象为包含诸多“要素”的一套知识本体，然后根据这些“要素”确定需要挖掘的数据源，将技术领域问题转化为机器学习的分类或回归问题，将技术领域背景知识储备转化为特征抽取的规则，构建出合适的训练语料；最后通过对模型的训练、验证，得到各种“要素”之间的相关关系，从而实现对未知问题的预测结果。仍以石墨烯技术领域为例，它包含的各种“要素”和“要素”之间的关系可以用图 1 进行展示。

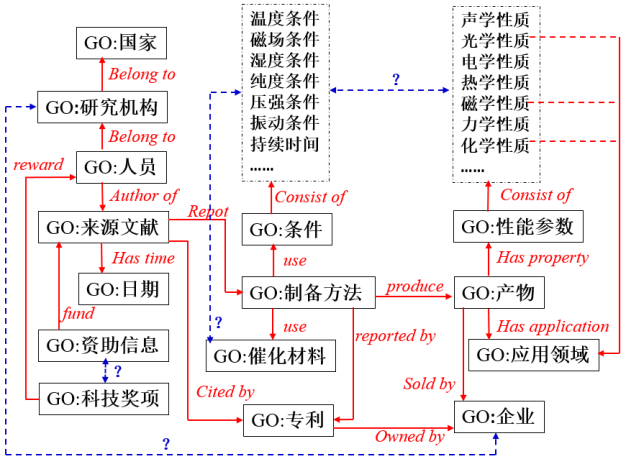

图 1 石墨烯技术领域知识本体（Graphene Ontology，GO）示意图

Fig.1 A diagram of Graphene Ontology

4.2 “情报认知模型库”的结构和功能

通过分析“情报认知模型”的要素可以看出，模型要素的来源既包括技术方法和性能参数细节，也包含文献来源信息。因此，在设计模型库的结构和功能

时，需要考虑文献库、知识库、算法库、实验数据/语料库等模块；为了方便模型库的建设和推广应用，可以分别设计“管理系统”“标引系统”“用户访问系统”，分别面向“模型库管理人员、建设和标引人员、用户”这 3 类人群提供模型库的访问途径。而在服务功能上，通过对模型库中的文本进行聚类分析、对各种性能参数等定量数据进行统计分析、对各种模型的效果和适用场景进行对比分析、对预标引数据和算法提供下载服务，可以实现研究热点识别、技术性能对比、科技领域智能分析方法推荐、知识发现算法辅助设计开发等功能，如图 2 所示。

值得指出的是，图 2 中并未显示模型库的不同模块、各模块的不同字段之间的相关关系，而这些相关关系是至关重要的。在未来的具体建设实施阶段，需要进一步设计各字段的数据类型/长度、录入必要性、字段录入的规范性等。

4.3 “情报认知模型库”的建设流程

考虑到“情报认知模型”的上述结构和功能特点，构建“情报认知模型库”的过程应当包括对模型的监测、收集、验证、标引、存储、分装、调用等，如图 3 所示。

（1）模型的收集和识别。在模型库的建设过程中，应当以来自科技决策层和科研一线、产业一线的情报需求为牵引，优先对具有重要应用价值的、有重复使用潜力的模型进行解析、验证和存储。应当尽量从经过同行评审的高质量学术期刊上搜集模型。另一方面，对模型的监测和积累可以嵌入科技情报工作者的日常工作业务中，即在周期性情报快报的监测和选题时、进行专题情报调研时，注意对情报人员扫描发现的有价值的情报认知模型加以关注，并将其纳入后续的验证环节。

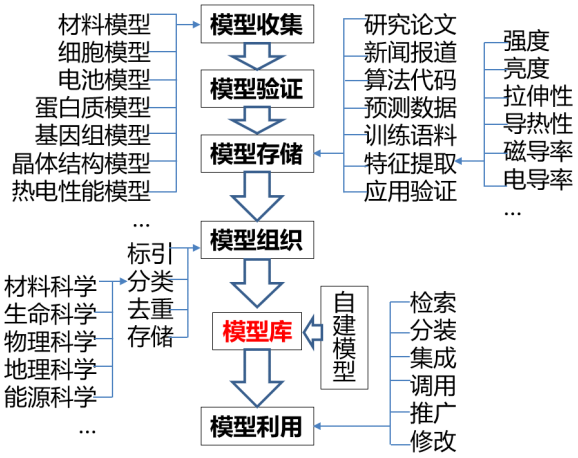

图 3　建设“情报认知模型”的工作流程

Fig.3 Workflow of constructing "intelligence cognition model database"

因此，可以纳入“模型库”的模型应当具备的标准包括：①权威性（由相关科学技术领域专业研究人员

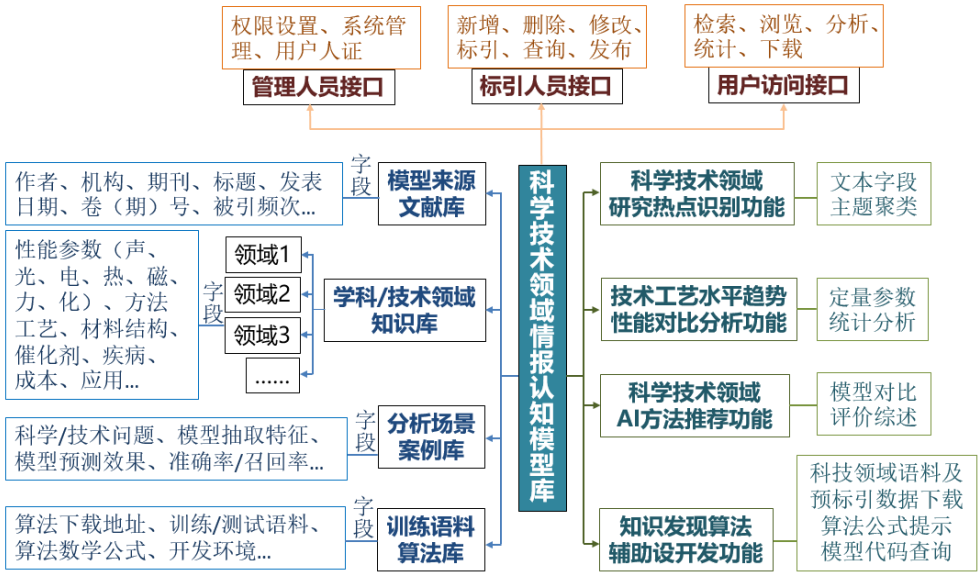

图 2　“情报认知模型库”的结构和功能

Fig.2 Structure and function of the "intelligence cognition model database"

DOI：10.13998/j.cnki.issn1002-1248.20-0969

开发，或发表在经过同行评议的高质量期刊上）；②完整性（包含对模型计算过程的定量描述、模型算法、输入和输出结果、训练语料、数据来源等）；③实用性（应当不是纯理论研究或推导，而是通过数据分析实现了知识发现，或确实解决了相关科学技术领域的实际问题）；④可重复性（模型的使用场景、步骤、条件清晰，可以由情报分析人员或相关领域科研人员对模型进行验证），等等。可以预见，随着模型库的建设，这些标准还会进一步充实和丰富。

（2）模型的验证。在初步发现有价值的情报认知模型后，对其的验证是至关重要的一步，这直接关系到建成后的模型库是否可靠、可信赖、可推广。在对“情报认知模型”进行验证时，首先应当对模型的水平、价值、应用范围做初步判断，对价值较低的模型进行剔除。对价值高的模型验证时，应当对模型的算法、代码、训练和测试语料数据进行下载，对研究论文中介绍的挖掘实验过程进行重现，对论文中的关键分析步骤和重要公示进行重点追溯，避免出现逻辑错误；在必要时可以聘请相关科学技术领域专家，对模型设计的学科背景知识和科学性、准确性进行把关。对模型的验证需要准备相应的计算环境和设备。如果模型验证需要的数据量和计算量过大，超出模型库建设单位的能力（例如一些采用大数据分析技术、或需要高性能计算设备的模型），那么可以暂时将“重现挖掘分析过程”省略；在将模型提供给服务对象（一般为具有相应计算环境和设备的单位）时，由服务对象完成对挖掘过程的重现验证。

（3）模型的标引。对“情报认知模型”的标引需要包括对报道和介绍模型的文献信息的标引，以及对模型自身特性的标引两个部分，如表 1 所示。

下面以香港大学 KOOHI-MOGHADAM 等 2019 年发表在期刊 *Nature Machine Intelligence* 上的论文 *Predicting Disease-Associated Mutation of Metal-Binding Sites in Proteins Using a Deep Learning Approach*[6]为例说明对模型标引的基本流程：这篇论文报道了使用深度学习方法预测蛋白质中金属结合位点的突变与疾病之间的相关性的一项研究。①在数据源方面，该研究首先从多个医学数据库下载了大量已知的金属结合位点蛋白质三维结构（来自 MetalPDB 数据库），以及人体细胞（致病/良性）突变数据（分别来自 ClinVar、

表 1　对“科学技术领域情报认知模型”的标引示例

Table 1　Indexing examples of the "scientific and technological intelligence cognition models"

科学技术领域	文献信息标引					模型信息标引				
	标题	时间	期刊	机构	作者	算法	数据源	抽取特征	预测问题	模型下载地址
生物	*Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach*	2019	*Nature Machine Intelligence*	香港大学	Koohi-Moghadam M, Wang H, Wang Y, et al.	多通道卷积神经网络（MCCNN）	MetalPDB（生物大分子金属位点数据库）、ClinVar（医学数据库）	金属结合位点的PDB结构、氨基酸序列、致病突变、氨基酸在蛋白质中的位置	金属种类与疾病种类之间的相关性	https://...
地质	*Hybrid computational intelligence models for groundwater potential mapping*	2019	*Catena*	Data & Analytics Practice, Virtusa, NJ, USA等	Binh Thai Pham, Abolfazl Jaafari, Indra Prakash,et al.	决策树分类器，基于旋转森林的决策树（RFDT），基于多重提升的决策树（MDT）等	印度中央地下水数据库（CGWB）	海拔，坡度，坡度，平面曲率，降雨，地形湿度指数（TWI），河流密度，岩性，土地利用和土壤	地下水井水面高度	https://...

Uniprot Humsavar 和 CancerResource2 数据库）；②在特征工程方面，该研究利用受控的医学主题词表对疾病名称进行了清洗，然后定义了 5 项数据特征（原始氨基酸类型、突变氨基酸类型、氨基酸在蛋白质中的位置、金属类型、相互作用类型），将蛋白质结构空间特征映射到三维点阵网格中，以矩阵形式在计算机中存储；③在机器学习模型的训练上，该研究将上述矩阵（网格）作为多通道卷积神经网络（MCCNN）的输入项，将不同金属的结合位点良性突变作为阴性训练集/测试集（输出项），而将致病突变作为阳性训练集/测试集（输出项）；④在分析效果上，该研究通过训练 MCCNN，对未知是否致病的突变情况进行了分类，最终发现了 1 256 种与疾病相关的错义突变，以及 261 种良性错义突变；在此基础上发现 10 种金属与 17 种疾病高度相关，例如锌结合位点的突变在乳腺、肝、肾、免疫系统和前列腺疾病中起主要作用，钙和镁的结合位点突变分别与肌肉疾病和免疫系统疾病有关，锰和铜结合位点突变与心血管疾病有关等。因此，这项研究中所蕴藏的“情报认知模型”的关键之处在于它在第②步（特征工程）中定义的 5 项“数据抽取特征”，以及后续对 MCCNN 的输入、输出、阳性/阴性训练语料的构建思路上；而相关算法、原始数据下载地址均可以重复利用，需要“模型库”构建人员进行标引和存储。最后这篇论文的标引结果如表 1 的第 1 行所示。

（4）模型的封装。对模型的标引是进行结构化存储、形成模型库的关键步骤。为了让模型库发挥科研基础设施和情报工具箱的效果，还可以将各种模型封装为可执行程序，方便科研人员和情报人员对模型的调用。

（5）模型的存储。“情报认知模型库”应该包含对报道和介绍各个模型的文献全文的存储、对模型算法和代码的存储、对模型训练和验证所采用的数据源和语料库的存储等。可以采用成熟的数据库构建方法对上述内容进行存储，并构建支持检索和调用的模型目录。

（6）模型库的服务和应用。“情报认知模型库”既可以供科技情报工作者查询和调用，也可以供相关科学技术领域的科研人员使用，同时也能够给研究机器学习算法和模型的技术人员以启发。在应用形式上，可以建设“情报认知模型库”门户网站，按照领域类别不同，对模型进行分类展示。科技情报人员可以在撰写情报报告时，通过门户网站查询、利用不同的情报认知模型，自动挖掘文献信息、生成情报观点，提高工作效率和分析水平。未来为了实现这一愿景，还需要进一步研究如何改进模型库的组织方式、提高模型库的自动化水平等。

5 结论和展望

构建“情报认知模型库”对开展科学技术领域情报工作具有重要意义。随着科研数据的爆炸式增长和信息技术的飞速发展，各种自动化、智能化的分析工具已经在科学研究、技术开发、科技情报工作中扮演至关重要的角色；“情报认知模型”中包含对各种科学技术领域知识的抽象、总结，能够发挥“从数据到知识”的重要桥梁作用，因此如果它们能被有效地收集、存储、封装，形成“情报认知模型库”，将具备广泛应用价值，发挥科技数据基础设施、科技情报分析工具箱的效果。

“对模型的验证”是建设“情报认知模型库”的关键步骤之一。为验证模型的可靠性、实用性，需要重现原始文献中所描述的利用该模型对科学技术领域数据进行挖掘、分析、实验、评估的过程。对于模型库的建设机构而言，如何在有限的硬件计算能力、技术分析能力条件下，重现一些涉及大数据分析任务的模型，可能是需要解决的难题之一。可能的处理方式包括采用“先存储，后验证”的方式，或寻求拥有相关技术条件、硬件资源的机构协助等。

“情报认知模型库”的构建需要情报人员、科学技术领域专家、信息技术人员的通力合作。任何一篇学术论文中设计的“情报认知模型”都是对纷繁复杂的自然现象和问题的抽象、简化、仿真，因此一定会存在信息损失，也一定是片面的；通过“情报认知模型”构建的科学技术领域知识本体，必须同该领域的专家智慧、研判相结合，才能保证整个科学技术领域的“知识地图”的完整性、合理性、权威性。

DOI：10.13998/j.cnki.issn1002-1248.20-0969

展望未来，在“情报认知模型库”初步建成后，如何对其进行推广服务，如何提升模型库检索、查询、调用的自动化水平，如何利用服务效果的反馈对模型库进行维护和更新升级等，还需要更进一步的研究。

参考文献：

[1] TONY H, STEWART T, KRISTIN T, et al. 第四范式: 数据密集型科学发现[M]. 北京: 科学出版社, 2012.
TONY H, STEWART T, KRISTIN T, et al. The Fourth Paradigm: Data-intensive Scientific Discovery [M]. Beijing: China science publishing & media ltd, 2012.

[2] 张霖. 关于数字孪生的冷思考及其背后的建模和仿真技术[J]. 系统仿真学报, 2020(4): 1-10.
ZHANG L. Cold thinking about digital twin and its modeling and simulation technology[J]. Journal of system simulation, 2020(4): 1-10.

[3] TSHITOYAN V, DAGDELEN J, WESTON L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature[J]. Nature, 2019, 571(7763): 95.

[4] ROSS D K. Automating chemistry and biology using robot scientists [EB/OL]. [2015-11-24]. http://ki2015.computational-logic.org/program/Keynote-Ross-King.pdf.

[5] SAID F, SAHAR V, CHRISTOPH L, et al. SEO: A scientific events data model [EB/OL]. [2019-03-24]. https://www.researchgate.net/publication/334285784_SEO_A_Scientific_Events_Data_Model?channel=doi&linkId=5d224819a6fdcc2462ca8858&showFulltext=true.

[6] KOOHI-MOGHADAM M, WANG H, WANG Y, et al. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach[J]. Nature machine intelligence, 2019, 1: 561-567.

[7] DATTILO A, VANDERBURG A, SHALLUE C J, et al. Identifying exoplanets with deep learning II: Two new Super-Earths uncovered by a neural network in k2 data[J]. The astronomical journal, 2019.

[8] WANG J, ZHANG Y, WEBBER S A H, et al. Solar flare predictive features derived from polarity inversion line masks in active regions using an unsupervised machine learning algorithm[J]. The astrophysical journal, 2020, 892(2): 140-149.

[9] PHAM B T, JAAFARI A, PRAKASH I, et al. Hybrid computational intelligence models for groundwater potential mapping[J]. Catena, 2019.

[10] 叶鹰, 武夷山. 情报学基础教程(第二版)[M]. 北京: 科学出版社, 2012.
YE Y, WU Y S. Basic course of information science (second edition)[M]. Beijing: China science publishing & media ltd, 2012.

[11] ESS-DIVE. ESS-DIVE Repository [EB/OL]. [2020-11-03]. http://ess-dive.lbl.gov/about/.

[12] MESRI. La feuille de route nationale des infrastructures de recherche [EB/OL]. [2018-05-06]. http://www.enseignementsup-recherche.gouv.fr/cid70554/la-feuille-de-route-nationale-des-infrastructures-de-recherche.html.

[13] DDBJ. Bioinformation and DDBJ center [EB/OL]. [2020-11-03]. https://www.ddbj.nig.ac.jp/index-e.html.

[14] ESRFI. Roadmap 2018: Strategy report on research infrastructures[EB/OL].[2018-05-06]. http://roadmap2018.esfri.eu/media/1060/esfri-roadmap-2018.pdf.

[15] 中华人民共和国科学技术部. 科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知[EB/OL]. [2019-06-05]. http://www.most.gov.cn/mostinfo/xinxifenlei/fgzc/gfxwj/gfxwj2019/201906/t20190610_147031.htm.
Ministry of Science and Technology of the People's Republic of China. Notice of the Ministry of science and technology and the Ministry of Finance on publishing the list of optimization and adjustment of national science and technology resource sharing service platform [EB/OL]. [2019-06-05]. http://www.most.gov.cn/mostinfo/xinxifenlei/fgzc/gfxwj/gfxwj2019/201906/t20190610_147031.htm.

[16] 中国科学院计算机网络信息中心. 中国科学院数据云[EB/OL]. [2020-11-03]. http://www.csdb.cn/pageAboutPlatform.
Computer network information center, Chinese academy of sciences. CAS data cloud [EB/OL]. [2020-11-03]. http://www.csdb.cn/pageAboutPlatform.

[17] 中国科学院文献情报中心. 科技文献大数据知识资源体系[EB/OL].[2020-11-03]. http://www.las.ac.cn/others/institute_characteristic.jsp.
National science library, Chinese academy of sciences. Big data knowledge resource system of science and technology literature[EB/OL].[2020-11-03]. http://www.las.ac.cn/others/institute_characteristic.jsp.